

CONDUCTING QUALITY IMPACT EVALUATIONS UNDER BUDGET, TIME AND DATA CONSTRAINTS



CONDUCTING QUALITY IMPACT EVALUATIONS UNDER BUDGET, TIME AND DATA CONSTRAINTS



Independent Evaluation Group,
The World Bank

Poverty Analysis, Monitoring and
Impact Evaluation Thematic Group
PREM Network
The World Bank



Acknowledgement

This booklet was prepared by Michael Bamberger, and was jointly sponsored by the World Bank's Independent Evaluation Group (IEG) and the Poverty Analysis, Monitoring and Impact Evaluation Thematic Group (PREM Network). It draws on earlier work commissioned by IEG — which was formerly known as the Operations Evaluation Department — and on a more detailed volume subsequently developed jointly by Michael Bamberger, Jim Rugh and Linda Mabry (2006). Valuable comments on earlier drafts of this booklet were provided by a number of Bank staff, including Judy Baker, Tara Bedi, Ariel Fiszbein, Emanuela Galasso and Howard White. The joint task managers of the booklet were Keith Mackay, Aline Coudouel and Markus Goldstein.

Klaus Tilmes
Manager
Knowledge Programs & Evaluation Capacity Development
Independent Evaluation Group

Copyright 2006
The International Bank for Reconstruction
and Development/The World Bank
1818 H Street, N.W.
Washington D.C. 20433, U.S.A.

All rights reserved.
Manufactured in the United States of America.
The opinions expressed in the report do not necessarily represent the views of the World Bank or its member governments. The World Bank does not guarantee the accuracy of the data included in this publication and accepts no responsibility whatsoever for any consequence of their use.

TABLE OF CONTENTS

1	Overview	1
2	Simplifying the evaluation design	5
	Case study 1: Reducing costs through a post-project cross-sectional design — gender and time-use impacts of the Ecuador cut-flower industry	9
3	Working with comparison group designs	10
	Case study 2: Using propensity scores to produce a matched comparison group — the Viet Nam Rural Roads Project	12
4	Working with secondary data	13
	Case study 3: Using multiple sources of secondary data — the Bangladesh Integrated Nutrition Project	15
5	Reconstructing baseline data	16
	Case study 4: Reconstructing baseline data — the Nicaragua Social Fund	18
6	Reducing the costs of data collection	19
	Case study 5: Using PRA techniques to reduce data collection costs — the Flores, Indonesia Village Water Supply and Sanitation Project	21
7	Strengthening evaluation designs when working under budget, time and data constraints	22
	Table 1. Design options for reducing data collection costs	8
	Table 2. How budget, time and data constraints affect the quality of an impact evaluation	23
	Endnotes	27
	References	29
	Additional Resources on Monitoring and Evaluation	31

1. Overview

Context and purpose of this publication

There is a growing appreciation within the development community of the merits of conducting evaluations. Evaluation is a powerful tool for learning about what works, what does not, and the reasons why. Evaluation is also an important accountability tool. For these reasons, the World Bank requires that all of its projects be evaluated. A growing number of developing countries also recognize the benefits from evaluation, and many are making efforts to institutionalize monitoring and evaluation systems as part of sound governance.¹

In this context, an increasing number of rigorous impact evaluations are being conducted, and it is important that the evaluation methods, findings and recommendations are as reliable as possible. At the same time, however, these evaluations can be expensive to conduct. Project and program managers who wish to conduct an evaluation are often faced with severe budget, time or data constraints — these can act as a disincentive to conduct rigorous evaluations. The purpose of this booklet is to provide advice to those planning an impact evaluation, so that they can select the most rigorous methods available within the constraints they face. The booklet is also intended to clarify the nature of the trade-offs between evaluation rigor and the budget, time and data which are available for an evaluation.² It is hoped that this booklet will encourage managers to conduct impact evaluations when they might otherwise have viewed them as too expensive or time-consuming to be conducted to a high standard. Thus the desired outcome is an increase in the quality and quantity of rigorous impact evaluations which are conducted.

An extensive literature is now available on appropriate methodologies for evaluating the impacts of development projects and programs. This booklet applies these methodologies to the real-world situations and constraints faced by task managers and researchers. It is intended to complement other recent World Bank publications including Baker (2000), Operations Evaluation Department (2004), Ravallion (2001, 2005), White (2006), and the methodological guidelines and impact evaluation case studies on the Bank's Poverty Impact Analysis, Monitoring and Evaluation website.³

Real-world evaluation constraints

Two sets of constraints shape the choice of evaluation methods. The first comprises budget, time and data constraints. **Budget constraints** affect the number of interviews that can be conducted, the ability to combine quantitative and qualitative data collection and analysis, the size and professional experience of the research team, and the analysis that can be conducted. **Time constraints** affect when the evaluation begins and ends, how long researchers can be in the field, and time available for feedback from stakeholders. When new surveys are conducted, **data constraints** affect the ability to collect information from a suitable comparison group and obtain baseline information on the project population; or to collect sensitive information and

to interview difficult-to-reach groups. When the evaluation is based on secondary data or when data are obtained through studies conducted by other agencies (piggy-back or synchronized studies), data constraints may affect the compatibility of sample coverage and timing, or whether the data cover the required variables and define them in the required manner.

In contrast, *program design and delivery mechanisms* determine how project participants are selected (randomly, according to administrative criteria, or through self-selection) and the quality and uniformity of project implementation. These design features, which derive from the inner working of a project, produce a range of project-specific issues which determine the range of evaluation methods which can be applied. These issues will not be covered here; readers interested in guidance on how to approach these questions of evaluation methods should see Ravallion (2001, 2005).

Real-world evaluation scenarios

We discuss two common scenarios. Under the first, the evaluator is involved from the start of the project but budget, time and data constraints or program design and delivery mechanisms limit the range of available evaluation designs. For example, it may not be possible to include a control group or to conduct a comprehensive baseline study on the project population, or there may be limits on the number of interviews that can be conducted.

Under the second scenario the evaluation is not commissioned until the project is nearing completion or has ended. Data may be collected in one of four ways (see White, 2006): from a project-specific survey; by piggy-backing a special module onto an ongoing survey; through a synchronized survey in which the project population is interviewed but a comparison group is obtained from another survey (national household survey, etc); or the evaluation is based on secondary data that include information on the project and potential comparison groups. A major constraint facing all post-project evaluation designs is the absence of a baseline study, and the options for addressing this are discussed in Section 5.

The requirements for a *quality impact evaluation* under real-world constraints

The challenge for the evaluator and the client is to decide whether it is possible to conduct a quality impact evaluation under the real-world constraints, and to select the strongest possible design within the particular set of budget, time and data constraints. For example: at what point does the sample become too small, or too limited in its coverage to permit sound statistical analysis? What are the criteria for assessing the adequacy of secondary data for estimating baseline conditions? And when is it possible to construct a valid counterfactual in the absence of a baseline study?

A quality impact evaluation must:

- Develop a set of indicators that can meaningfully and reliably define and measure project inputs, implementation processes, outputs, intended outcomes and impacts.
- Develop a logically sound *counterfactual* presenting a plausible argument that observed

changes in outcome indicators after the project intervention are in fact due to the project and not to other unrelated factors, such as improvements in the local economy or programs organized by other agencies.

- Determine, in accordance with accepted statistical procedures, whether a project has contributed to the intended impacts and benefited a significant proportion of the target population.

In addition many evaluations are required to:

- Assess the distribution of benefits among different sectors of the target population.
- Identify factors influencing the magnitude and distribution of the impacts.
- Assess the sustainability of impacts over time.

When resources are not a constraint, the conventional evaluation approach is to use a pre- and post-intervention project and control group comparison similar to the following:

Figure 1. Pre-test—Post-test Control Group Design			
Time	T ₁ Start of project (baseline)	T ₂ Project intervention (this may last a few weeks or as long as several years)	T ₃ End of project
Project group	P ₁	X	P ₂
Randomized or non-randomized control group	C ₁		C ₂

The pre-test—post-test control group comparison represents the **counterfactual**—what would have happened to the project population if the project had not taken place.

There are a number of methodological advantages from evaluation designs in which subjects are randomly assigned to the project and control groups as this avoids systematic differences between the two groups prior to the project. However, in many operational settings random assignment is not possible so the two groups will be matched as closely as possible using procedures such as matching on *observables* or *propensity scores* (see Section 3). Both groups are surveyed at the start of the project (T₁) and again after project implementation (T₃). If the groups are well matched, then any statistically significant difference between the two groups on impact variables in T₃ is indicative of a potential project impact. However, the differences could also be partly explained by different experiences of the two groups during the project implementation period, (e.g., an unrelated project that targets only the control group).

Two important design elements for the estimation of project impacts are: a carefully selected control (or comparison) group and pre- and post-intervention comparison of the two groups. The first element is essential for the formulation of a logically sound counterfactual. While the second element, observations at two points in time, is usually desirable, it is possible to use post-project project and control group cross-section data only. If the cross-sectional data set has the observable characteristics that are used to assign the program to beneficiaries, and we are confident that unobservable characteristics do not play a role in program participation, then

propensity score matching may provide an acceptable lower cost option. While a limitation of propensity scores is that they are not able to control for all pre-existing differences between the two groups, they are usually the best option where baseline data are not available — as well as being much cheaper. In practical terms, the more we understand about the project context and the participant selection process, the more confidence we can have in the validity of propensity score matching.

Many of the real-world scenarios discussed in later sections concern situations in which one or more of these four observation points (P_1 , P_2 , C_1 , and C_2), has been eliminated — either as a deliberate strategy to reduce cost and time, or because circumstances did not permit the collection of data at every point (for example, when the evaluation did not start until late in the project or when the available budget did not permit collection of baseline data on a control group). A central question we will address is: how do evaluation approaches under cost and time constraints affect the validity of the evaluation design and the conclusions derived from the analysis? What are acceptable compromises that permit valid findings and what are the minimum methodological requirements without which a study can no longer be considered a quality impact evaluation?

Section 2 considers a range of options for simplifying the design of an evaluation, and examines the implications for evaluation rigor and cost. Section 3 presents options for selecting comparison groups, while Section 4 looks at the use of secondary data. Strategies for reconstructing baseline data, to improve an evaluation's rigor, are examined in Section 5, while options to reduce the costs of data collection are outlined in Section 6. Finally, the ways in which budget, time and data constraints reduce the rigor of an evaluation are reviewed in Section 7. This section also presents a range of options for addressing each of these constraints.

2. Simplifying the Evaluation Design

Randomized designs in real-world project settings

In randomized evaluation designs, subjects (individuals, communities, schools, health clinics, etc) are randomly assigned to the project and control groups, ensuring the two groups have the same distribution of observed and unobserved characteristics at the start of the project. This ensures that post-intervention differences in impact scores are not due to initial differences (selection bias) in the characteristics of the two groups. Despite the potential benefits, random assignment has only been used in a small proportion of development evaluations for one of the following reasons: target communities, organizations or individuals are selected according to certain administrative or political criteria (the poorest or most needy or locations where the project is most likely to succeed), subjects are self-selected (individuals or groups make the decision to participate), or political considerations make randomization impractical. Under both of these selection scenarios — random and non-random selection — there are likely to be systematic differences between the project and non-project groups with respect to factors affecting impacts. For example, people applying to village banks for small business loans may already have the self-confidence or experience to launch a successful business.

There may be a significant number of cases in which randomization is not only feasible, but the best way to do things: for example, pilot projects, projects where resources are limited relative to demand, or where the efficacy of the treatment is unknown. However, for all of the reasons discussed above, randomization is not an option for most development interventions and, given these constraints, most of the discussion that follows concerns cases in which a program is not administered randomly and where randomized evaluation designs are not an option.

Gain scores are defined as:

the pre-project/post-project difference in scores on the impact indicator (*single difference*). When a control group is used the *gain score* is the difference in the pre-project/post-project change for the project and comparison groups (*double difference*).

Factors explaining project and control group differences in randomized and non-randomized evaluation designs

Evaluation designs in which a separate sampling procedure must be used for the non-intervention group are referred to as *non-equivalent control group* or *comparison group* designs. Technically, the term *control group* should only be used when there is random allocation. The factors that may contribute to post-project differences in indicators of project impact between project and non-project groups (*gain scores*) in true experimental designs, randomized designs in field settings, and non-equivalent control group designs can be summarized as follows:

- In true experimental designs, the conditions of the two groups are carefully controlled during the treatment, and gain scores should only be attributable to the project effect.

- For randomized designs in real-world scenarios, it is rarely possible to control the project setting. Consequently, gain scores may be related to *differential time varying effects* such as contamination/spillovers affecting the control group, different patterns of attrition, and the influence of selection or non-selection on the behavior of subjects and stakeholders.⁴ For example, in an evaluation of the impact of flip-charts on academic performance in schools in Kenya, it was found that parent-teacher associations became more active in schools receiving flip charts, and parents were more likely to encourage their children to study, so that some of the changes in academic performance may not have been directly related to the pedagogical value of the flip charts. In other cases, the presence of a donor funded project may encourage government agencies to provide additional support (not included in the project design) as they would like the project to succeed, while in other cases communities or organizations not selected may become demoralized or government support may be less forthcoming.
- For non-equivalent control group designs, gain scores may also be influenced by sample selection bias with respect to characteristics captured in the survey (*observables*) — these can be controlled in the analysis — and to characteristics not captured in the survey (omitted variables or *unobservables*) that cannot be controlled in the analysis. While some of the omitted variables could be easily included in future surveys, others such as motivation or intelligence would be extremely difficult to capture.

Reducing costs and time by simplifying the evaluation design

As data collection often represents more than half of the cost of an evaluation (Baker, 2000), this section discusses ways to reduce the costs of data collection by simplifying the evaluation design. Table 1 estimates that for evaluations requiring the collection of survey or other kinds of primary data (rather than analyzing existing survey and other sources of secondary data), some of the simplified evaluation designs can reduce the costs of data collection by more than 50 percent. However, there are important trade-offs between cost-cutting strategies, the quality of the evaluation and the validity of the conclusions. In fact some of the most economical designs (Designs 5 and 6 discussed below) can no longer be considered quality impact evaluations, although they may produce operationally useful findings.

Robust and less-robust designs available in real-world contexts

We use as our reference point the *robust general purpose impact evaluation design* described in Figure 1 — this is Design 1. All the simplified designs described in this section eliminate one or more of the following observation points: the baseline (pre-test) comparison group, the baseline project (treatment) group or the post-test comparison group. It is not possible to eliminate the post-test project group as this is always needed to measure the project effect. The five simplified design options (see Table 1) are:

- Design 2: *Delayed pre-test/post-test comparison group design*. Similar to Design 1, except that the evaluation does not begin until the project has been underway for some time, usually as

part of the mid-term project review. If the project is delayed, this design may not be much weaker than Design 1 but if the project starts on time it is significantly weaker. Modest cost savings will result from a shorter consultant contract period.

- Design 3: *Pre-and post-intervention project group and post-intervention comparison group*. In this design, there is no pre-intervention comparison group. While methodologically weaker than Design 1, this design can ensure a reasonable degree of analytical rigor. The potential 25% data collection cost-saving results from the elimination of the baseline comparison group.
- Design 4: *Post-intervention project and comparison groups with no baseline data*. This widely-used design defines the post-intervention comparison group as the counterfactual, assuming that any observed differences between the two groups (after controlling for observable household characteristics) are due to the effects of the project and not to any unobservable pre-intervention differences between the two groups (see Case Study 1 below). Possible data collection cost savings of up to 50% can be achieved from eliminating all baseline surveys.

Two common evaluation designs that do not qualify as sound *impact* evaluation designs

The two following designs, while not qualifying as sound impact evaluation designs (see Section 1), are included because of their popularity, and because if used with appropriate caveats can potentially provide some insight into project effects.

- Design 5: *Pre- and post-intervention project group comparison (reflexive comparison)*. This design eliminates the comparison group and consequently does not provide a logically sound counterfactual. The design identifies the project impacts only under very strong (and usually improbable) assumptions that there were no time-dependent changes at play. This design is widely used both because the elimination of the comparison group can cut data collection costs by up to 50% , and because there are many situations in which data are available for the project group (usually from project surveys and administrative records) but not for a comparison group.
- Design 6: *Post-intervention project group without baseline data or a comparison group*. This is the weakest design and although it is widely used to estimate project effects⁵ it cannot be considered as producing rigorous quantitative estimates of project impact. Frequently this design is used when the evaluator is operating on an extremely tight budget (sometimes as little as \$10,000) and may only be able to spend a few weeks in the field. Estimates of change (impacts) are based on a combination of qualitative data such as recall, key informants, focus groups and participatory group techniques such as participatory rapid assessment (PRA), project records and secondary data from public service agencies (e.g., ministry of health or education), censuses and other government data. Secondary data are only used to obtain global comparisons of the project communities with similar areas, and not for household-level analysis. Depending on the nature of the design, costs can be reduced by 75% or more.

Table 1. Design options for reducing data collection costs^(a)						
Evaluation Design	Base-line	Treat-ment	Mid-term	Post inter-vention	% cost savings compared to Design 1	Comments
The reference design: A robust general purpose impact evaluation design						
1. Pre- and post intervention project and control group design with randomized or non-randomized assignment.	P ₁ C ₁	X		P ₂ C ₂	0	Strongest design in most real-world cases where the evaluation starts at the same time as the project.
Progressively less robust designs						
2. Delayed pre-test/post-test design with a comparison group. Evaluation does not start until around project mid-term.		X	P ₁ C ₁	P ₂ C ₂	0	Useful design when evaluation starts at mid-term. If implementation is delayed this may not be much weaker than Design 1. Possible modest cost savings due to shorter evaluation contract period.
3. Pre- and post-intervention project group with post-intervention only comparison group.	P ₁	X		P ₂ C ₂	25	While weaker than the previous designs this is relatively strong as it permits comparison over time and post-project transversal analysis.
4. Post-implementation only project and comparison groups with no baseline data.				P ₁ C ₁	50	A common design when the evaluation begins late in the project cycle or when the project has ended. The lack of baseline data makes it difficult to control for initial differences between the two groups, but this can be partially resolved with propensity scores. However, using propensity scores may require a larger sample.
Two frequently used designs where the lack of an acceptable counterfactual almost always disqualifies them as quality impact designs						
5. Pre- and post-implementation project group analysis with no comparison group.	P ₁			P ₂	50	A common design when data are only collected on the project group. Methodologically weak as using P1 as the counterfactual requires heroic assumptions about time-varying effects and individual unobservable variables.
6. Post-implementation project group with no baseline data and no comparison group.				P ₁	75-90	The weakest design but one which is commonly used when evaluations have to be conducted late in the project with very limited time and money. Qualitative methods, project records and aggregate secondary data are used to estimate the counterfactual.
Key to symbols: P = randomized or non-randomized selection project group. C = randomized or non-randomized (non-equivalent) selection of control/ comparison group. P ₁ , P ₂ , C ₁ , C ₂ indicate the first and second administration of the survey instrument to the project and comparison groups respectively.						
^(a) The cost savings are based on the assumption that surveys or other primary data collection will be required. Estimated reductions in data collection costs, compared to Design 1, are based on the number of the 4 data collection points (pre and post intervention, and project and comparison group) that are eliminated. In cases where piggy-back or synchronized surveys are used, or the evaluation relies on secondary data, the cost saving can be greater.						

Case Study 1: Reducing evaluation costs through a post-project cross-sectional design — gender and time-use impacts of the Ecuador cut-flower industry

The evaluation illustrates how the costs of data collection can be significantly reduced through a post-intervention cross-sectional design (Design 4) in which data are only collected at one point in time after the intervention is already operating. The weakness of the design is that it does not determine to what extent assumed project impacts are in fact due to pre-existing differences between the groups not captured in the survey (unobservables).

The purpose of the study was to assess the impacts of women's employment on the allocation of paid and unpaid labor within the household. The study compares household labor allocation in the geographical areas with access to the cut-flower industry, which offers women much higher wages, with the situation in similar areas which do not have access to this source of employment. The study uses a post-intervention cross-sectional design in which socio-economic interviews were conducted in May-June 1999 with a sample of 562 households during which observations were obtained on all 2567 household members over the age of 10. The sample included "treatment" households living in the valley where the cut-flower industry operated and "control" households living in a similar valley at a distance of some 200 km with no access to the cut-flower industry. A detailed accounting of time use was obtained both for a 24-hour period and also for the previous week. Multiple regression analysis, with a dummy variable for employment in the cut flower industry, was used to control for household characteristics that might influence the dependent variable (the amount of time that men and women devote to housework).

The study found that when the wife works the husband spends more time on domestic activities and that the increase was greatest when the wife worked in the cut-flower industry where her earnings relative to her husband's were greatest.

While regression analysis improves the match of post-intervention project and control households, the weakness of this design is that it cannot determine whether the observed time-use differences already existed before the cut-flower factories opened. This is a serious weakness of the evaluation design as it is possible that one of the factors determining the location of the flower industry might have been that it was known women in this area already had a high labor force participation rate. This problem could have been addressed by using some of the techniques discussed in this booklet for reconstructing baseline data (see Section 5).

Source: Newman, 2001.

3. Working with Comparison Group Designs

Project participants are selected in one of three ways:

- randomly from among all members of the target group (eligible individuals, communities, schools, etc);
- using administrative or political criteria (the poorest families, technical feasibility to provide access to existing infrastructure, groups considered most likely to succeed, etc);
- or subjects themselves elect to participate (self-selection).

For the purposes of impact evaluation, randomized participant selection has the advantage that the distribution of observed and unobserved characteristics can be assumed to be similar for the project and control groups so that post-intervention gain scores are not determined by initial differences between the two groups, and consequently are likely to be due to the project intervention.

However, for the reasons discussed in Section 2, random assignment is practiced in only a small proportion of development projects, so that for most impact evaluations a **quasi-experimental design** must be used in which different sampling procedures are used for the selection of the project and comparison groups. This has important implications for the analysis of project impacts, as post-intervention gain scores may be due to sampling bias (differences in the characteristics of the two groups), rather than to the effects of the project. This section discusses issues and approaches in the use of quasi-experimental designs for the evaluation of project impact.

Options for selecting comparison groups under budget, time and data constraints

Matching areas on observables. The researcher, in consultation with clients and other knowledgeable persons, identifies characteristics on which comparison and project areas should be matched (e.g., access to services, type or quality of house construction, economic level, central location or remoteness, types of agricultural production). The researcher then combines information from maps (and sometimes Geographic Information System (GIS) data and aerial photographs), observation, secondary data (censuses, household surveys, school records, etc) and key informants to select comparison areas with the best match of characteristics. When operating under real-world constraints it will often be necessary to rely on easily observable or identifiable characteristics such as types of housing and infrastructure. While this may expedite matters, it is important to keep in mind the potential for unobservable differences, to address these as far as possible through qualitative research, and to attach the appropriate caveats to the results.

Matching individuals or households on observables. Similar procedures are used to match individuals and households. Sample selection can sometimes draw on previous or ongoing household surveys, but in many cases researchers must develop their own ways to select the sample. Sometimes the selection is based on observable physical characteristics (type of housing, distance from water and other services, type of crops or area cultivated) while in other cases selection is based on characteristics that require screening interviews, such as economic status,

labor market activity, or school attendance. In these latter cases the interviewer is given quotas of subjects with different characteristics to be located and interviewed (quota sampling).

Pipeline sampling: The comparison group is defined as individuals, households or communities selected to participate in the project but who have not yet done so (see Ravallion, 2005 and White, 2006). Often large projects such as housing or community infrastructure are introduced in phases over several years and some beneficiaries will not begin to receive services until several years after the start of phase one. When there are no major differences between the characteristics of families or communities scheduled for each phase, the later phases can provide a good comparison group for the earlier phases. These procedures are also economical to use. However, project design and selection criteria must be carefully reviewed because there will often be systematic differences between the phases. For example, phase one may start with the poorest families or alternatively with the more centrally located or better-off areas, and in both of these cases the characteristics of communities in later phases are likely to be different.⁶

Regression discontinuity designs: In cases where a program is assigned using a clear threshold for eligibility comprised of one or more criteria, this program assignment rule can be used for evaluation. The basic idea is to compare individuals, communities or units just above the threshold and hence not eligible for the project (the comparison group) with those just below the threshold who are eligible (the treatment group). This procedure requires that the treatment rule is fairly enforced in practice, and the selection criteria are not subject to manipulation by potential beneficiaries.

Propensity score matching. Statistical matching procedures can be used when secondary data are available to select subjects, communities or sites (schools, clinics, etc) with similar characteristics to the project group. The most common matching method is **propensity score matching** where each project participant is statistically matched to a group of non-participants (nearest neighbors) on a set of relevant characteristics. The mean value of the outcome variable is calculated for the nearest neighbors, and this is compared with the outcome score for the project participant to estimate the *gain score* (see Ravallion, 2006 and Baker, 2000 for a summary of propensity score matching).⁷ An issue when working under budget constraints is that the use of propensity scores will often require a larger sample to ensure that the best matching variables are identified.

Multiple comparison group designs. When projects are implemented in different ways, or participants receive different combinations of services, it may be possible to use different comparison groups for different treatments.

Potential biases and issues when using comparison groups

Comparison groups are subject to a number of potential biases or problems. Many of these may become more problematic when evaluations are conducted under real-world constraints. Some of the common issues include:

- Projects target *all* communities or subjects with particular characteristics (the poorest families, the largest slums, the communities or individuals thought most likely to succeed),

making it difficult to find close matches for a comparison group. Where there are clear rules for project selection, it may be possible to use regression discontinuity to compare project areas with other locations. However, the situation sometimes arises, for example, where almost all of the poorest urban families live in a few very large slums, while most slightly better-off families live in very different kinds of communities so the match is not very close.

- When individuals or communities self-select into the program, it is difficult to identify factors determining the decision to participate and consequently to find a good comparison. This is a serious analytical challenge as those who elect to participate may be those most likely to succeed (e.g., women who apply to small-business development programs may already have entrepreneurial experience) and consequently positive outcomes could be more related to participant characteristics than to the project.

Project and comparison groups may differ in terms of factors not covered in the survey (*omitted variables or unobservables*). Sometimes these factors could be easily included in future studies, while in other cases the evaluator may not be aware of some important factors, or they may be difficult to measure (for example, reasons for choosing to participate in a project).

Section 7 describes a number of strategies for addressing common problems affecting non-equivalent control group designs and ways to strengthen them. This covers: sample selection bias, sample size issues, post-selection bias, inconsistencies in project implementation, unreliability of outcomes measures, contextual influences on project implementation and outcomes, rapid assessment studies and triangulation.

Case Study 2: Using propensity scores to select a matched comparison group — the Viet Nam Rural Roads Project

The survey sample included 100 project communes and 100 non-project communes in the same districts. Using the same districts simplified survey logistics and reduced costs, but communes were still far enough apart to avoid “contamination” (control areas being affected by the project). A logit model of the probability of participating in the project was used to calculate the propensity score for each project and non-project commune. Comparison *communes* were then selected with *propensity scores* similar to the project communes. The evaluation was also able to draw on commune-level data collected for administrative purposes that cover infrastructure, employment, education, health care, agriculture and community organization. These data will be used for contextual analysis and to construct commune-level indicators of welfare and to test program impacts over time. The administrative data will also be used to model the process of project selection and to assess whether there are any selection biases.

Source: Van De Walle and Cratty, 2005.

4. Working with Secondary Data

We indicated in Section 1 that data may be collected in one of four ways (see White, 2006):

- from a project-specific survey;
- by piggy-backing a special module onto an ongoing survey;
- through a synchronized survey in which the project population is interviewed but a control group is obtained from another survey (national household survey, etc);
- or the evaluation is based on secondary data collected for some other purpose but that include information on the project and potential control groups.

Almost all evaluations will draw on secondary data, even when surveys are conducted, and in many cases secondary data will be the main or only source of information. Consequently, for most evaluations the question is not whether to use secondary data but rather how to ensure their adequacy and quality for a particular evaluation.

The advantages of secondary data

Secondary data can be a useful way to reduce costs and to save time. When post-intervention project/comparison group designs are used, secondary data often provide the only way to *reconstruct* baseline conditions of the project and comparison groups prior to the start of the project. For this, and most other designs, they can be used to strengthen the *counterfactual* estimate of what would have been the situation of the project population if the project had not taken place.

Some of the most common types of secondary data include:

- National census data
- General household surveys such as Living Standards Measurement Surveys (LSMS).
- Specialized surveys such as Demographic and Health Surveys (DHS).
- Administrative data collected by line ministries and other public agencies (school enrolment, use of health facilities, market prices for agricultural produce).
- Studies conducted by donor agencies, non-government organizations and universities.
- Administrative data from the project agency or ministry.
- Mass media (newspapers, television documentaries, etc). These can be useful, among other things, for understanding the local economic and political context of each project location

Another important application for secondary data is through meta-analysis in which the impacts of comparable projects or interventions in this or other countries provide estimates of the size of the effects that could be expected from a well-designed project. Meta-analysis can be particularly useful in estimating the required sample sizes for the project and comparison groups as (other things being equal) the smaller the expected effect size, the lower the power of the test and the larger the required sample to detect project impacts when they do exist.⁸

While secondary data are extremely valuable for evaluation, the information was likely collected for a different purpose and data sources must be carefully assessed before being used. Some of the potential issues that must be considered when using secondary data include (see Bamberger, Rugh and Mabry, 2006, Chapter 5):

- Time differences between the start of the project (when baseline data are required) and the time when secondary data were collected or reported.
- How closely does the sample approximate the target population? (e.g., does the survey cover private as well as state schools? Does it cover informal as well as formal sector employment? Does it cover both men and women as well as other groups of interest such as the elderly?)
- Was information collected on all key project variables and outcome indicators and are the data adequate for the purposes of the evaluation? Often one or two simple proxy indicators have to be used to measure complex outcome indicators (for example, using indicators of health services delivered as a proxy for health impacts, or using the volume and types of vehicles, and the number of new businesses, as indicators of the impacts of rural roads).

It is also important to assess the quality and completeness of the information. When information is collected for administrative purposes, there may be no quality control and the information may be incomplete, inaccurate or biased (as when schools have an incentive to inflate enrolment rates or test scores, or the police may under-report crimes). This is particularly important in the case of impact evaluation as the incentive to misreport is higher if the service delivery unit knows these data will be used for an evaluation.

Using secondary data to address budget, time and data constraints

The following are some of the common ways in which secondary data are used to reduce time or costs or to reconstruct baseline conditions or comparison groups:

- Project administrative data are used as a proxy baseline. For example, applicants for low-cost housing often have to provide information on their current housing; applicants for micro-credits provide information on their current economic activities and income; and infrastructure planning studies collect information on the accessibility and quality of current infrastructure.
- Household survey data are used to match project populations with similar non-project communities or households which will be used as a comparison group.
- Census and household survey data can be used to construct a comparison group.
- Records from schools or local health centers in comparison areas not affected by the project are used to estimate the counterfactual for programs affecting health and education. These records may also be used to compare utilization of new project-supported facilities with traditional schools, clinics, etc. in project areas. This may be important in order to control for the fact that users may switch to the upgraded facilities, and these may not be comparable to the average user at a non-program facility.
- Two or three different secondary sources can be used to create separate comparison groups for consistency or for different types of analysis.

Case Study 3: Using multiple sources of secondary data — the Bangladesh Integrated Nutrition Project

The purpose of the Operations Evaluation Department (OED) evaluation was to assess the impact of the Community Based Nutrition Component of the Bangladesh Integrated Nutrition Project (BINP) on reducing severe malnutrition among children and pregnant women. The evaluation used three separate secondary survey data sources: the BINP project evaluation, which collected data at baseline, midterm and endline; the survey carried out by Save the Children at the end of the project; and the Nutritional Surveillance Project (NSP) carried out by Helen Keller International. Between them, these surveys covered not only nutritional outcomes but also a wide range of process indicators, allowing the application of a theory-based approach. The findings from the evaluations were contradictory, and there were questions about the appropriateness of the comparison groups. In addition, the OED study conducted a meta-analysis of studies in similar countries to assess what percentage of reductions in mortality per 1000 live-births could reasonably be expected from the hiring and training of traditional birth attendants. A re-analysis of the BINP data using NSP data to construct a better-matched comparison group through propensity score matching found that both *height for weight* and *weight for age* indicators had improved more in project areas.

The combination of these different secondary data sources permitted the construction of a causal chain, and this identified several missing links that helped explain why the project impact, although statistically significant, was disappointingly low in operational terms (only around 5%).

Several lessons were learned on the use of secondary data for evaluating nutritional projects. Different sources of rich secondary data are sometimes available, and when combined can improve estimates of project impact by strengthening comparison groups. These data sources can also help describe the causal chain through which project impacts are to be achieved. However, the study also showed that there are often gaps in the information with respect to project administration and implementation, and also much less data were available on demand side factors.

Source: White, 2006.

5. Reconstructing Baseline Data

Post-test only designs are weakened by a lack of baseline data

Evaluations not commissioned until a project is nearing completion do not have the option of collecting baseline data. While statistical techniques such as **propensity scores** (see Section 3) can improve the project and comparison group matching, post-implementation evaluations are usually forced to define the post-implementation comparison group as the *counterfactual*, and these designs cannot control for *unobserved* differences between the two groups at the time of project launch (see Case Study 1). This is particularly problematic when participants are *self-selected*, as individuals or communities that elect to participate in a project are often those most likely to succeed (e.g., female micro-loan recipients who already have entrepreneurial experience, self-confidence and control over household resources). Consequently, the lack of baseline data makes it difficult to separate project effects from pre-existing differences.

Sometimes, earlier surveys or census data can provide estimates of pre-intervention conditions (see Section 3), or other kinds of secondary data such as project records, records from local schools or health facilities, or from other government agencies, may provide general estimates of baseline conditions (see Section 4) but usually without household level comparisons. However, when no secondary data are readily available, the researcher may consider some of the qualitative techniques described below.

An assessment of strategies for reconstructing baseline data

Secondary data from national household surveys, censuses and similar studies, conducted around the time of project launch may identify the project population and include relevant data on the project population and/or potential comparison groups. Sometimes the surveys provide good data on the comparison group but the number of project participants is too small to permit detailed analysis.

- It is important to assess how well each data set satisfies the needs of the evaluation. For example, the data may not have been collected at exactly the right time, they may not cover all of the project or comparison population, they may not include all of the critical information required for the evaluation, or there may be questions about the reliability of the data. However, experience from World Bank evaluations shows that it is often possible to find high quality secondary data covering most of the baseline information needs.

Administrative records collected by the project (e.g., the characteristics of families or communities applying for housing, school places, small business loans or basic services) often provide baseline data on project participants, but usually not comparison groups.

- Data must be checked for quality, completeness and consistency of reporting. Agencies may introduce reporting biases if it is known that the records will be used to evaluate project performance. Another problem is linking administrative records to post-project data

as subject identification numbers are often not complete or reliable. Without subject matching, the evaluation cannot use panel studies — which are better for controlling for individual or community (time-invariant) *unobservables*.

Records from schools, clinics, savings and loan cooperatives, sales from local agricultural markets, etc., can sometimes provide baseline comparison group data.

- There are sometimes questions about the completeness of the records or about systematic biases (e.g., teachers may have an incentive to over-report the number of children attending schools or the police may under-report crime).

Recall: Respondents can be asked to recall their earlier situation with respect to school attendance, use of health facilities, or time and cost of travel.⁹

- Recall data are subject to potential biases due to problems of memory, under- or over-reporting, and the distortion of socially desirable or undesirable behavior. Results are also very sensitive to the time period covered and how questions are formulated. Except in a few areas such as expenditures and fertility behavior, where systematic research has been conducted on recall, there are usually very few empirical studies estimating the magnitude or direction of bias (see Bamberger, Rugh and Mabry, 2006, pp. 97-99). This may argue for the comparison of multiple measures of the impact variable.

Participatory Rapid Assessment (PRA) is now used as a generic term for a range of participatory techniques in which communities, rather than individuals, report on community conditions, problems and changes over time (Kumar, 2002). Community groups can provide estimates of, for example, the volume and quality of water; crop production and sales; travel time and costs; and time use. Several independent sources of information should always be triangulated (see Case Study 5) — triangulation involves the systematic comparison of estimates obtained from two or more independent sources.

- In addition to problems of recall bias, PRA also faces problems of representativity (who attends the group discussions?), and group dynamics (do certain groups dominate the discussions?). The collection of group data also changes the unit of observation at which impact is measured, reduces the sample size, and presents difficulties of integrating community-level findings with individual and household-level survey data.¹⁰

Key informants can be interviewed about the situation prior to the project. The views of several different independent informants should be triangulated as the views of any individual informant are likely to include (intentional or unintentional) bias. While key informants only provide global information at the community or group level, this can be useful where survey data are not available. This approach can also be used to assess some of the important *unobservables* such as, for example, unique characteristics of project participants or communities which make them more likely to be successful in the project.

Administrative data collected by the project can be adapted to include questions that will prove useful for a later impact evaluation.¹¹

Strengthening the quality and analysis of reconstructed baseline data

- The checklist in Section 4 can be used to assess the weaknesses of all potential secondary data sources and Section 7 suggests ways to address some of the common threats to validity of the different approaches described in this section.
- Triangulation should be used to check secondary sources. It can be used to compare estimates from two or more questions included in a survey, or to compare qualitative estimates of, for example, economic status, labor force participation, or school enrolment survey data. Potential triangulation sources include: direct observation; other secondary sources; key informants; stakeholder surveys; PRA; photographs and newspaper articles. These techniques will help the researcher understand potential *unobservables* or other sources of bias in the data.

Case Study 4: Reconstructing baseline data: the Nicaragua Social fund

This case illustrates the use of four secondary data sources to produce independent estimates of accessibility and impacts of the Nicaraguan Emergency Social Investment Fund (FIS) and to reconstruct baseline conditions, and the need for independent impact estimates when project communities and beneficiaries are not selected randomly.

FIS provides latrines, schools, health posts, and water supply to low-income communities selected from among those applying to participate in the program. Selection criteria gave priority to poor communities but also took into account community capacity to implement the project. Communities are not selected randomly, nor can they be defined simply as the poorest. Several sources of secondary data were used to create independent comparison groups:

- The 1998 LSMS national survey produced poverty maps to guide the selection of FIS intervention areas and to identify project and non-project communities for each project component (water supply and sanitation, health, education, etc). The samples were then combined with administrative data on project selection criteria to calculate **propensity scores** estimating the probability of living in an area of influence of the respective FIS component. The project gain scores (school enrolment, age for grade, repetition, etc) were estimated as the average difference between the project and matched comparisons for each impact variable.
- The census was combined with administrative data to select *choice based samples* of direct beneficiaries and potential beneficiaries because the FIS sample did not include enough households for statistical comparisons.

The LSMS and FIS household studies were conducted when the FIS had already been operating for up to 5 years in some communities and no baseline study had been conducted. The evaluation mainly relied on a post-test comparison of treatment and non-treatment areas using propensity scores to match the two groups. The sources described above were used to strengthen the analysis through the reconstruction of baseline data. Recall was also used to obtain information on pre-project measures of project outcome variables.

Source: Pradhan and Rawlings, 2000.

6. Reducing the Costs of Data Collection

Reducing the length or complexity of the survey instrument

Consultations with the client to identify what information is essential and what is only “interesting” can often produce significant reductions in the length or complexity of the survey instrument and consequently the costs and time required for data collection.

Reducing the number of interviews conducted

As data collection can often represent more than half of the evaluation budget (Baker, 2000), reducing the sample size can produce significant cost savings. However, there are trade-offs as smaller samples reduce statistical precision of estimates and the level of disaggregation of the analysis. Some of the key determinants of the required sample size for a particular evaluation include: the estimated average treatment or effect size, the desired power of the statistical test, the mean and variance of the underlying variables, the required level of statistical precision, whether or not a comparison group is used, the types of disaggregated analysis and whether a one or two tailed statistical test is required.¹² Based on these considerations, the following are some of the options for reducing the required number of interviews:

- Accepting a lower level of statistical precision (e.g., 90% confidence interval instead of 95%) or a lower *statistical power of the test* (e.g., accepting a 20% instead of a 10% risk of rejecting a real project impact).¹³ Of course, this increases the chance of wrongly judging whether a project has, or has not, had an impact.
- Reducing the level of statistical disaggregation of the analysis (e.g., only obtaining results for the total project population rather than comparing impacts on different groups or the effectiveness of different components of the project).
- The larger the expected effect size, the smaller the sample required to find a statistically significant impact.

Replacing individual interviews with community-level data collection

While individual interviews provide the most detailed and statistically precise information, there are other, more economical ways to collect comparable, although not as quantitatively precise, community or group-level information:

- An observation checklist can be used at the community level to estimate, for example, travel and transport patterns, time spent in collecting water and fuel, or service quality and utilization. Observations are collected at different locations in the community (roads and footpaths, routes to water sources, or at local clinics). While easy and economical to administer, observation checklists usually reduce the number of observations from a large number of individuals to a small number of communities. The question must then be addressed of whether the number of communities must be increased to find statistically significant results, and if so whether the overall cost savings will still be obtained. Ideally, in order to strengthen the analysis, additional information should be collected on related attributes (covariates) at the household or community levels, which might further reduce the cost savings.

- Focus groups, PRA techniques or community interviews can also be used to obtain community level estimates of service utilization and quality, agricultural production, time use or gender division of labor. The issue of the consequence of reducing the number of observations — discussed above — again arises.

Reducing interview costs

- For evaluations of health or education programs, nurses or teachers who know the field and have social acceptability can be recruited to conduct the interviews and possibly help with data coding. For more general socio-economic surveys it may also be possible to recruit university students or even high-school graduates. While the payment per interview will be lower than hiring a professional survey company, additional money and time must be budgeted for training and possibly for a higher level of supervision.
- Using self-administered questionnaires rather than having the information collected by a team of enumerators. This option is, of course, only available when surveying literate populations such as evaluations of secondary education programs. The approach may also introduce bias in response rates, including systematic difference in bias across treatment and comparison groups. Biases are harder to control as the enumerator is usually not present when the survey is being completed.

Using diaries to reduce the cost of collecting income and expenditure data for an impact evaluation

In an evaluation of the impacts of investments in housing on the ability of poor households to cover basic expenditures, 100 families agreed to keep daily records of all income and expenditure over the period of one year. The only compensation was the families were able to choose a small household gift each month. This proved to be much more economical than having a team of enumerators visit each family at least once a week. However, the success of this approach depended on the unusually high level of cooperation of the families all of whom were participating in a self-help housing project.

Source: Valadez and Bamberger, 1994.

Electronic data collection

While many electronic technologies are not available in many developing countries where evaluations are conducted, these technologies are rapidly becoming more widely accessible, and there are possibilities for significantly reducing the costs and time of data collection:

- Replace face to face surveys with telephone interviews. Even when not all respondents have personal phones, many can take calls in a community center or a friend's house. With the falling price of mobile phones, respondents can also be given or loaned a mobile phone.

Telephone interviews can save money and time

In an evaluation of a school voucher program in Colombia, most of the interviews were conducted by phone. As access to a phone was a criterion for participation, phone interviews did not introduce a significant sample selection bias and certainly saved money and time.

Source: Angrist et al., 2002

- E-mail surveys are becoming increasingly used at the organizational level. Many respondents, even very poor ones, now have access to a community telecenter or internet café, and so e-mail may start to be used more widely in some types of surveys.
- Automatic counters can be used to record, for example, the number of people entering a building, or pedestrian and vehicular traffic.
- GIS systems and aerial photographs can sometimes be economical ways to obtain physical information on, for example: the number and size of houses and other forms of construction, transport patterns, and cultivated land.

Cost sharing

It is sometimes possible to share the cost of data collection and analysis with another agency, or to “piggy-back” the study (particularly when the instrument is short) on another survey. Sometimes a module can be administered to a sub-sample of households covered in an ongoing national household sample survey.

Case Study 5: Using PRA techniques to reduce data collection costs — the Flores, Indonesia Village Water Supply and Sanitation Project

This case study illustrates the use of PRA techniques as a cost-effective way to reconstruct baseline conditions and to assess changes in access and quality of services. The sample of communities is sufficiently large to permit statistical analysis.

The evaluation assessed the effects of a community management approach to the provision of water supply and sanitation on access, effective use and sustainability of the services. The Methodology for Participatory Assessment (MPA), a form of PRA, was used to obtain the perspective of different community groups as well as local and national government agencies. A stratified random sample of 63 sites was drawn from the universe of 260 sites where village water supply and sanitation projects were being carried out by a number of different public and NGO agencies in addition to the World Bank project. The following participatory techniques were used: (a) household welfare classification in which households were classified in an open community meeting into better-off, worse-off and in-between; (b) social mapping and transect walks in which male and female representatives of the 3 strata assessed access, quality and sustainability of the services; (c) committee interviews using various card-sorting techniques to produce ordinal assessments of the management and quality of the services. The findings were presented and discussed in open community meetings. Ordinal ratings of the different aspects of the program were compared with ratings of the degree of community participation in planning and management and the extent to which women were involved. While the evaluation design made it possible to compare water projects implemented by a number of different agencies, it did not include a comparison group of villages without any project and consequently the evaluation only compares the effectiveness of the different service delivery systems and does not permit a with/without project comparison. There was also no discussion of selection bias due to the special characteristics of the communities selected for the projects. The evaluation cost \$150,000 (including \$45,000 for international consultants) and was completed in 12 months.

Source: Hopkins and Mukherjee, 2005.

7. Strengthening Evaluation Designs When Working Under Budget, Time and Data Constraints

The effects of budget, time and data constraints on the quality of impact evaluations

Conducting impact evaluations under budget, time or data constraints increases the difficulty of dealing with four sets of threats to the quality of the design and the validity of the conclusions.¹⁴ While these four sets of *threats to the validity of evaluation conclusions* can affect all evaluations, they are more difficult to manage when working under real-world constraints. Table 2 describes common problems under each of these four categories and identifies which constraints tend to contribute to each problem:

- *Threats to overall quality of the evaluation design and implementation:* resource constraints may limit attention to the design of the evaluation, instrument development and testing, and client consultations; and there may be pressures to cut data collection costs by hiring cheaper interviewers, or reducing interviewer training and supervision. It may also be more difficult to use mixed-method approaches and triangulation for quality control and to fully check on the adequacy of secondary data sources.
- *Threats to statistical analysis:* The constraints make it harder to take measures to strengthen the sampling frame or address sample selection bias. There may also be pressures to reduce the number of data collection points (e.g., eliminating baseline data or comparison groups). There may also be pressures to reduce sample size, which reduces the power of the statistical test and limits possibilities for disaggregated analysis.
- *Theoretical coherence and adequacy of the counterfactual:* the constraints make it difficult to conduct the exploratory studies, client consultations and workshops needed to develop a program theory explaining how the program is expected to achieve its objectives and how the magnitude and distribution of impacts is affected by contextual variables and the project implementation process. The constraints also weaken the counterfactual by eliminating data collection points or reducing access to secondary data that can strengthen the comparison group.
- *Generalizability of findings:* When the evaluation cannot control for sample bias or analyze contextual factors affecting outcomes in specific locations, this increases the risk of coming to wrong conclusions about whether the project could be replicated.

Ways to address the effects of budget, time and data constraints on the validity of the evaluation conclusions

The previous paragraph describes how budget, time and data constraints exacerbate many of the most common threats to validity of evaluation designs. When resources are constrained, trade-offs will almost always have to be made on how resources will be used. Some of these involve saving money by, for example, hiring cheaper interviewers or reducing the number and depth of case studies; while others involve decisions on whether to, for example, invest scarce resources to increase sample size, improve the coverage and quality of the sampling frame or reduce non-response rates by requiring more revisits to households. Table 2 is meant to provide a basic, indicative checklist to guide the reader in potential dimensions in which tradeoffs may need to occur (for more see Bamberger, Rugh and Mabry, 2006). While these tradeoffs may be inevitable, framing them in the context of what dimension they are likely to compromise the results is a useful way to inform your decision. Nonetheless, there may be ways to lessen the constraints, and there also may be things which should never be compromised. The following section provides some guidance on these.

Table 2. How budget, time and data constraints affect the quality of an impact evaluation			
Problems (threats to validity) caused by different constraints)	Constraints contributing to each problem		
	Budget	Time	Data
A. Problems affecting overall quality of evaluation design (threats to internal validity)			
Insufficient attention to planning, client consultation and developing rapport with local consultants	✓	✓	
Insufficient attention to instrument development and testing	✓	✓	
Less time for follow-up on evaluation findings	✓	✓	
Exclusion of difficult-to-reach groups and difficult-to-obtain information	✓	✓	✓
Less application of mixed-method approaches, so that triangulation consistency checks cannot be used	✓	✓	
Pressures to find cheaper interviewers and less resources for training and supervision	✓	✓	
More reliance on rapid qualitative methods	✓	✓	
Harder to check on the adequacy of secondary data	✓	✓	✓
B. Problems affecting sample design and statistical analysis (threats to statistical validity)			
Less opportunity to apply mixed-method approaches	✓	✓	✓
Less resources to improve quality of sampling frame	✓	✓	
Harder to address sample bias and improve matching	✓	✓	✓
Poorer quality of sample implementation	✓	✓	✓
Smaller sample size — risk of false negatives	✓		
Pressures to eliminate collection of project or control group baseline data or post-intervention comparison group	✓		✓
Less disaggregated analysis	✓	✓	✓
C. Problems affecting the coherence of theory and the validity of the counterfactual (threats to construct validity)			
Less time and resources to develop a program theory model so that key concepts and indicators may be less well defined and key hypotheses may not be identified or may be wrongly specified	✓	✓	✓
Less use of multi-method approaches and triangulation	✓	✓	✓
Weaker (smaller or not as well matched) control/comparison group	✓		✓
Weaker or no baseline data	✓	✓	✓
D. Problems affecting the generalizability of findings and recommendations concerning the replicability of the project in other settings or with different groups (threats to external validity)			
Lack of attention to sample bias	✓		✓
Weak analysis of contextual factors contributing to success or failure in specific locations	✓	✓	✓

Strengthening the overall quality of the evaluation design

- Even when operating under budget and time constraints, it is always important to allow sufficient time to meet with clients and key stakeholders to understand *their* information needs, deadlines and constraints. Any design decisions on ways to save money or time must be made in consultation with clients so that they fully understand and accept the trade-offs involved.
- When time is the main constraint, it is possible to get off to a quick start by organizing video conferences with local agencies and researchers and by commissioning preparatory studies to be completed *before* the arrival of foreign consultants.
- Consider the cost implications of how selection bias and instrumental variables are addressed. Working with project staff to address selection bias during project design will often be more economical than conducting costly surveys during the post-implementation evaluation to match participants and comparison groups on a set of difficult-to-identify selection criteria. Project managers may agree to introduce more explicit participant selection criteria to avoid some biases, or administrative data collection may be strengthened during participant selection so that actual selection criteria will be better documented. Applying and documenting clear selection and rejection criteria strengthens the possible types of analysis.
- Developing a program theory (i.e., logic) model and articulating the *effects chain* through which impacts are expected to be achieved can help identify the critical assumptions and issues on which the scarce evaluation resources should be focused.
- Peer review, ideally a standard component of any evaluation, is very helpful when assessing the threats to validity of measures to address real-world constraints, by bringing another perspective to bear on the tradeoffs being made.
- When the budget does not permit the use of a household sample survey, PRA and other qualitative techniques can often provide community level estimates of, for example, water consumption or use of new sanitary facilities (see Case Study 5). However, if these techniques are to be used for impact evaluation a sufficiently large sample of communities will be required to permit statistical analysis and it will be necessary to assess whether there are still cost savings.
- When operating with smaller than ideal sample sizes, it is sometimes possible to develop cost-effective mixed-method approaches that strengthen validity by providing two or more independent estimates of key output, outcome or impact indicators. For example, if key informant estimates of changes in quantity and reliability of water supply are consistent with household surveys, this can increase the credibility of the estimates. However, the use of mixed-methods increases data collection costs so it is important to determine whether this strategy does contribute to overall cost reduction. It is also important to remember that if the sample is too small, it will still not be possible to apply statistical tests for hypotheses testing — even if the mixed-methods increase the plausibility of a causal relationship.

Strengthening the sample design and statistical analysis

- *Sample size issues:* Reducing sample size is a tempting way to save money, but smaller samples increase the risk of *false negatives* (wrongly assuming the project did not have an impact). Statistical power analysis (see Section 4) is a useful way to ensure the proposed sample will be large enough for the purposes of the analysis.

Strengthening the theoretical framework and the validity of the counterfactual

- Rapid assessment studies are a cost-effective way to develop the program theory models discussed earlier.
- When working with small samples, mixed method approaches can provide a cost-effective way to understand key concepts and to improve their measurement.
- The counterfactual can be strengthened by reconstructing baseline conditions (even when using propensity score matching designs that rely on post-test comparisons without baseline data) and where necessary strengthening comparison groups using the techniques discussed in Section 5.
- Ensure time and resources are allocated to develop program theory. Simpler models can be developed in a relatively short period of time. Often the evaluator will find that although program staff all have their own ideas of the program objectives and how they will be achieved there is no official formulation of the program model and this has to be elicited during interviews, workshops and review of program documents. It is, however, possible to do this in a cost-effective and rapid manner.

Strengthening the generalizability of conclusions

- Use rapid assessment methods (key informants, focus groups, observation, etc) to identify the similarities and differences between the project and comparison groups and to understand how this might affect generalizability of findings. Also use mixed methods to interpret the results — good qualitative work can shed light on the process that got you the measured outcome, even if you did no process analysis initially.
- Use contextual analysis (qualitative or quantitative description and analysis of the local economic, political, institutional and socio-cultural context in each project location) to understand how local factors might affect outcomes and to assess how far the operation of these factors is unique to this particular context and how far they could be generalized.¹⁵
- Multivariate analysis should be used where possible to strengthen the match of project and comparison groups and hence strengthen the validity of projections of the conditions under which the project could be replicated.

Assessing when it is feasible to conduct an impact evaluation

Before implementing the evaluation, an *evaluability assessment* should be conducted to determine whether the requirements for conducting a quality impact evaluation (see Section 1) can be met with the available resources, timeline and data availability. This is conducted after all possible avenues have been explored for strengthening the evaluation design, using the techniques

discussed earlier, and the best available design has been proposed. If an acceptable *impact* evaluation cannot be conducted within these constraints, the resources and time frame must be renegotiated, the scope and objectives of the evaluation revised, or the evaluation cancelled. Remember it is often possible to conduct an operationally useful assessment of potential project effects even when conditions do not permit a rigorous impact evaluation.

ENDNOTES

¹ The World Bank, with the support of the Independent Evaluation Group, is helping a growing number of governments to develop and strengthen their monitoring and evaluation (M&E) systems.

² A number of case studies have been chosen to illustrate the use of certain methods, but this does not necessarily mean that they are impact evaluations in the sense that we use that term here.

³ <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/0,,contentMDK:20381417~menuPK:773951~pagePK:64165401~piPK:64165026~theSitePK:469372,00.html>

⁴ If these contextual factors have a consistently different effect on project and comparison groups, this may artificially increase or reduce the average project effect.

⁵ CARE International recently estimated that about 50 percent of all of its project evaluations conducted worldwide are forced by budget, time, data and logistical constraints to use some variant of this design (see Bamberger, Rugh and Mabry, 2006, Chapter 10).

⁶ The World Bank's Operations Evaluation Department tried to use a pipeline comparison group to evaluate an irrigation project in Andhra Pradesh, India but found that farmers covered by the later phases were typically more remote and different in other ways from phase one farmers. (See White, 2006, p. 14).

⁷ See Baker, 2000, Box 3.1 for a brief explanation of propensity scores and Ravallion, 2005, for a more detailed discussion.

⁸ For example, the Bangladesh Integrated Nutrition Project conducted an international meta-analysis to estimate the number of deaths per 1,000 live births that it would be reasonable to expect could be saved through recruiting traditional birth attendants. It was found that the range was 5-7 deaths averted per 1,000 live births. This was used to confirm that the project's target of 7 deaths averted per 1,000 live births was realistic (see White, 2006, case study 3). The same analysis could have been used to estimate expected effect size when estimating the required sample size for an impact evaluation.

⁹ For example, the evaluation of the Eritrea Community Development Fund used a post-test evaluation design. In order to reconstruct baseline conditions prior to the construction of community schools families were asked to recall for the period before the schools were built, which of their children attended school (particularly gender differences in attendance), how long it took children to travel to school outside the village and the cost of travel. Similar questions were asked with respect to the village health centers. Family responses were cross-checked with information from key informants (village elders, teachers, local government officials, etc). The evaluators believed that recall was relatively reliable in this case because the construction of a school or clinic in a remote village was an event that everyone could easily remember and there was no obvious reason why respondents might wish to distort the information. A similar recall technique might be more difficult to apply in an urban area where families often have access to several different school or health facilities and where other programs may have been introduced over the same period — making it more difficult for respondents to focus on the particular project of interest to the evaluation.

¹⁰ The LSMS survey package includes a module for community and service level analysis (Frankenberg, 2000, Volume 1, pp. 315-338) and there is experience in the treatment of these multilevel analytical issues.

¹¹ For example, applicants for a low-cost housing program are usually asked where they currently live as well as to provide information on their current income. It would be relatively easy and economical to request more detailed information on their current living conditions (size and quality of the dwelling unit, access to services, transport facilities), economic conditions and labor market activities. However, depending on the work pressure of the reception staff and the number of applicants, it might be necessary to hire additional staff to collect this information — so the activity would not be completely costless.

¹² For a discussion of the effect of each of these factors on sample size see Bamberger, Rugh and Mabry, 2006, Chapter 14.

¹³ The Statistical Power of the Test is the probability of wrongly rejecting a statistically significant association between the project and the dependent variable (impact indicator). The risk of wrongly rejecting a significant impact can be reduced by increasing the power of the test (from, for example, the conventional 0.8 level, which accepts a 20% risk of wrongly rejecting the impact to, for example, Power = 0.9 which reduces the risk to 10%). Increasing Power increases the sample size, while accepting a lower Power can significantly reduce the sample size. (See Bamberger, Rugh and Mabry, 2006, Chapter 14).

¹⁴ These categories are based on the four sets of threats to validity of evaluation conclusions discussed in the quasi-experimental design literature. The threats are: internal conclusion validity (overall quality of the evaluation design); statistical conclusion validity; construct validity (coherence of theory and adequacy of counterfactual); and external validity (generalizability of findings). For a fuller discussion of threats to validity, together with checklists and worksheets for applying the concepts in the field, see Bamberger, Rugh and Mabry, 2006, Chapter 7 and Appendices 1, 2 and 3.

¹⁵ For a discussion of contextual analysis see Bamberger, Rugh and Mabry, 2006, Chapter 9.

REFERENCES

- Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King and Michael Kremer. 2002. 'Vouchers for private schooling in Colombia: evidence from a randomized natural experiment'. *American Economic Review* 92 (5): 1535-1558.
- Baker, Judy. 2000. *Evaluating the Impacts of Development Projects on Poverty: A Handbook for Practitioners*. Washington, D.C.: The World Bank.
- Bamberger, Michael, Jim Rugh and Linda Mabry. 2006. *Real World Evaluation: Working under Budget, Time, Data and Political Constraints*. Thousand Oaks, CA: Sage.
- Frankenberg, Elizabeth. 2000. 'Community and price data', in Margaret Grosh and Paul Glewwe (eds.) *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standards Measurement Study*, Chapter 9. Washington, D.C.: The World Bank.
- Hopkins, Richard and Nilanjana Mukherjee. 2005. 'Assessing the effectiveness of water and sanitation interventions in villages in Flores, Indonesia', in Operations Evaluation Department, 2005, *Influential Evaluations*, 22-30. Washington, D.C.: The World Bank.
- Kumar, Somesh. 2002. *Methods for Community Participation*. London: ITDG Publications.
- Newman, Constance. 2001. *Gender, Time Use, and Change: Impacts of Agricultural Export Employment in Ecuador*. Policy Research Report on Gender and Development. Working Paper Series No. 18. Washington, D.C.: The World Bank.
- Operations Evaluation Department (OED). 2004. *Monitoring and Evaluations: Some Tools, Methods and Approaches*. Washington, D.C.: The World Bank.
- _____. 2004. *Influential Evaluations: Evaluations that Improved Performance and Impacts of Development Programs*. Washington, D.C.: The World Bank.
- _____. 2005. *Influential Evaluations: Detailed Case Studies*. Washington, D.C.: The World Bank.
- Howard White. 2006. *Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank*. Washington, D.C.: The World Bank (forthcoming).
- Pradhan, Menno and Laura Rawlings. 2000. 'The impact and targeting of social infrastructure investments: lessons from the Nicaraguan Social Fund'. *World Bank Economic Review* 16 (2): 275-295.
- Ravallion, Martin. 2001. 'The Mystery of the Vanishing Benefits: An introduction to Impact Evaluation,' *World Bank Economic Review* 15 (1): 115-140.
- _____. 2005. *Evaluating Anti-Poverty Programs*. Policy Research Working Paper No. 3625. Washington, D.C.: The World Bank.
- _____. 2006. *Evaluating anti-poverty programs. Handbook for Agricultural Economics* (edited by Robert Evenson and Paul Schulz), Volume 4. North-Holland.

Valadez, Joseph and Michael Bamberger. 1994. *Monitoring and Evaluating Social Programs in Developing Countries*. Washington, D.C.: The World Bank.

Van De Walle, Dominique and Dorothyjean Cratty. 2005. *Do Donors Get What They Paid For? Micro Evidence on the Fungibility of Development Project Aid*. World Bank Policy Research Working Paper No. 3542. Washington, D.C.: The World Bank.

Additional Resources on Monitoring and Evaluation

World Wide Web Sites

- World Bank Independent Evaluation Group:
<http://www.worldbank.org/ieg/>
- World Bank Independent Evaluation Group — impact evaluation:
<http://www.worldbank.org/ieg/ie/>
- World Bank — impact evaluation:
<http://www.worldbank.org/impacetevaluation/>
- Building government monitoring and evaluation systems:
<http://www.worldbank.org/ieg/ecd/>
- Monitoring and Evaluation News:
<http://www.mande.co.uk/>